Volume 12 Issue 01 Published, February 21, 2008 ISSN 1535-864X DOI: 10.1535/itj.1201.06

Intel® Technology Journal

Technology with the Environment in Mind

Dynamic Data Center Power Management: Trends, Issues, and Solutions

Dynamic Data Center Power Management: Trends, Issues, and Solutions

David Filani, Digital Enterprise Group, Intel Corporation
Jackson He, Digital Enterprise Group, Intel Corporation
Sam Gao, Digital Enterprise Group, Intel Corporation
Murali Rajappa, Digital Enterprise Group, Intel Corporation
Anil Kumar, Digital Enterprise Group, Intel Corporation
Pinkesh Shah, Digital Enterprise Group, Intel Corporation
Ram Nagappan, Digital Enterprise Group, Intel Corporation

Keywords: dynamic power management, data center, data center operations, energy efficiency, platform management policy, power limiting, power capping

ABSTRACT

In this paper we examine the challenges of increasing data center power consumption and higher energy costs in the face of ever-increasing computing needs. An examination of how power is allocated to computing resources in data centers shows that current methods do not result in optimal use of available data center power and space. We identify requirements that server platforms must address to solve data center power problems, and we offer a solution that includes a platform resident Policy Manager (PM). The PM monitors power and thermal sensors and enforces platform power and thermal policies. We explain how the PM can be used as the basis of a data center power management solution. We present results from a Proof of Concept (PoC) implementation, and we conclude by showing that a policy-based approach is powerful for maximizing power allocation within a given power envelope and increasing server density in data centers.

INTRODUCTION

One of the biggest challenges for data center operators today is the increasing cost of power and cooling as a portion of the total cost of operations. As shown in Figure 1, over the past decade, the cost of power and cooling has increased 400%, and these costs are expected to continue to rise. In some cases, power costs account for 40-50% of the total data center operation budget. To make matters worse, there is still a need to deploy more servers to support new business solutions (Figure 2). Data centers are therefore faced with the twin problem of how to deploy new services in the face of rising power and cooling costs. In a recent survey of data centers (Figure 3), 59% identify power and cooling as the key factors limiting server deployment.

If these trends continue, the ability of data centers to deploy new services will be severely constrained. To overcome this constraint, data centers have three choices: expand power and cooling capacity, build new data centers, or employ a power management solution that maximizes the usage of existing capacity. The first two choices can be very expensive because they involve capital expenditure for purchasing and installing expensive new power delivery equipment. For this reason, the power management approach bears close examination, and this approach is the focus of the rest of our paper. For previous work in this area, the reader is referred to Felter et al. [4] who examine the benefits of dynamic power budget allocation, Femal [5] who examines the benefits of monitoring and coordinating power distribution to achieve higher application throughput, [6] where a framework to monitor power is discussed, and Bianchini [7] who presents a survey of energy management techniques by type of server system.

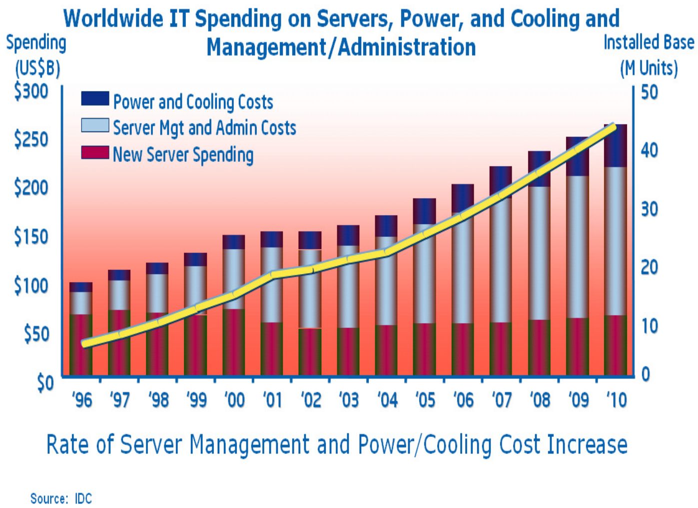

Figure 1: IDC Report of data center cost structure and trend

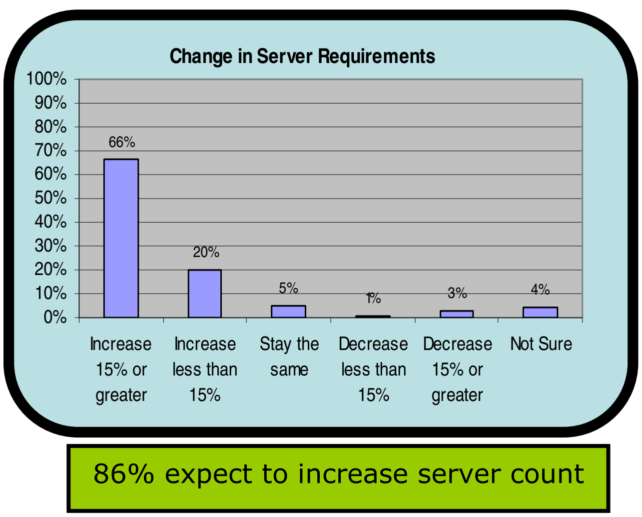

Figure 2: Expected growth in server count

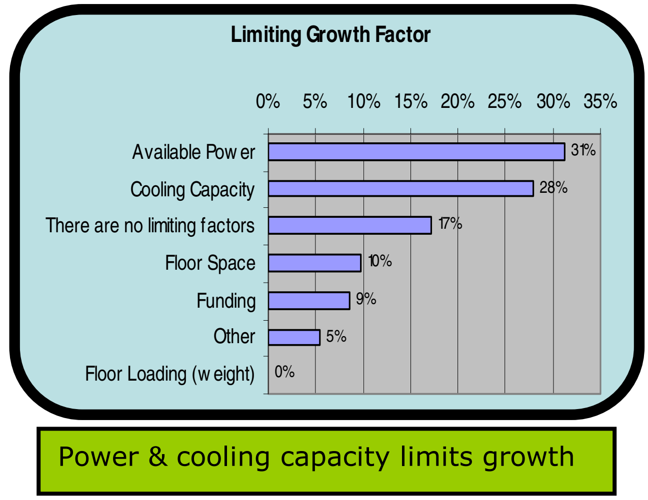

Figure 3: Factors limiting server growth

Our paper is organized as follows. In the next section we describe the current data center power allocation approach and the resulting problems; we follow this by proposing a new method for dealing with data center power allocation and describe the resulting server requirements; we then describe the role and functions of a platform resident Policy Manager (PM) and show how it addresses these requirements. Finally, we present the results of a PM Proof of Concept (PoC) and the benefits it offers data centers.

CURRENT POWER ALLOCATION METHODS

A typical data center power distribution hierarchy is designed to deliver a fixed amount of power to the room and then to each rack. The challenge of the data center operator is to determine the number of servers for each rack while ensuring that the overall rack (hence room) power consumption does not exceed the limit. To do this, the operator must make certain assumptions about the maximum power consumption per server. For most data centers, there are two ways of determining this: 1) using server nameplate power value, and 2) using a derated nameplate value.

The server nameplate value, which is marked on the server by the manufacturer, is the maximum possible power value that the server can consume. Actual power consumption is typically much less that the nameplate power. Most data center operators are aware that typical server power consumption never reaches the nameplate value, and one way for them to increase server density is to derate the nameplate power by a certain percentage—depending on the workload that is deployed on the server.

While derating, as opposed to using nameplate value, can improve server density, it is obvious that both methods are not optimal. The result is the following:

1. Under-utilization of available power for computing: A static allocation of power based on the worst-case scenario planning leads to inefficiencies and does not maximize the use of available power capacity.
2. Under-population of rack space: This is a direct result of the static allocation approach described above. The wastage of rack space is more severe when racks are populated using the nameplate power.
3. The two factors above result in higher energy costs than actually necessary, primarily from the cost of cooling the room that is sparsely populated with servers.
4. Unnecessary power cooling capacity expansion: Although existing power capacity is not being utilized effectively (as described in 1, 2, and 3 above), if the data center needs to deploy new services, the operator has no choice but to expand power/cooling capacity or even build a new data center at very high costs.

From the analysis above, we can see that the current power allocation approach results in wastage and high total cost of ownership. A better approach is needed.

DYNAMIC POWER MANAGEMENT APPROACH

The fundamental problem with the current approaches just described is that they are not based on measurements of actual power consumption. The data center operator has no visibility into how much power each server (and hence the rack and room) is consuming at any given time, neither does he/she understand the power consumption pattern over time. Without that visibility, the operator (or data center management software) is unable to decide how much power to allocate to servers/racks based on actual need. A better approach is to allocate power and populate racks using the following steps:

1. Monitor actual power consumption to understand average and peak power utilization for the server/rack.
2. Dynamically allocate power to groups of servers to maximize power/space utilization while staying within the power constraints determined by the power and cooling capacities.
3. Dynamically reallocate power when necessary to accommodate shifts in power needs of servers.

To implement this dynamic power allocation approach, server platforms must address key power management requirements.

POWER MANAGEMENT REQUIREMENTS FOR SERVER PLATFORMS

In this section, we identify the server requirements that address the issues of data center power and cooling efficiency, power allocation, and power provisioning.

Power Measurement Requirements

Actual power consumption for each server must be measurable at any point in time. This allows a power-monitoring module to collect power consumption data for each server and aggregate power-usage values at the rack and data center levels over a period of time.

Real-time power monitoring will allow data center managers to see the trend of power usage over time. This allows her to identify key values such as minimum, typical, and peak usage. This information is very useful in planning for future expansions, in identifying where there might be power and cooling constraints, and in locating areas where new servers can be deployed without violating power and cooling constraints. In addition, the minimum, typical, and peak values can be used to determine an appropriate power policy for each rack and hence each server in the rack. System power can be monitored by communicating with power supplies that support the Power Management Bus (PMBus) interface [13].

It is also desirable to monitor the power consumed by server subsystems: CPU, memory, fans, disks, etc. As shown in [8] an understanding of subsystem power consumption characterization for workloads can be valuable for power adaptation. Hence subsystem power monitored values can be used for intelligent fine-grained power control and optimization algorithms.

Power monitoring eliminates the usage of nameplate or derated power value to determine the number of servers for each rack. Visibility into actual power consumption values allows IT personnel to determine the optimal number of servers to deploy per rack.

Power Control Requirements

As mentioned above, the ability to monitor server power consumption is in itself useful for power and cooling capacity planning purposes. But another key reason for monitoring power consumption is to determine appropriate power policies that can be set for servers in a data center. Such policies can be enforced autonomously

in the platform, by following the platform as a service model described in [2].

One important power policy is power capping. In their study of power usage at a large Internet services data center at Google, Fan et al. [1] found that power capping offers two advantages: first, it acts as a safety valve by protecting the power distribution hierarchy against overdraw; and secondly, it enables effective usage of the available power, thereby increasing rack population. As such, dynamic power capping is a primary power control requirement that must be addressed by a power management solution.

Power Usage Reporting Requirements

To monitor power and dynamically allocate/reallocate power in a data center, a standard interface must exist between data center management software and the server being managed to do the following:

- Monitor actual power consumption over a planning period to understand historical usage patterns for capacity planning.
- Provide current as well as peak power, minimum power, and average power over an interval.
- Notify the higher-level management system if the power policy cannot be enforced.
- Send alerts to higher-level management if a certain power threshold is reached.

Regulation Requirements

The American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) publish a guideline for providing the data needed for designing and provisioning a data center [9, 10]. A PM's power usage reporting capability should support this requirement. A PM can be an instrumental piece of monitoring carbon credit/generation. As carbon generation may have caps and limits in the future, a PM will be a fundamental tool to monitor and track power use, which is correlated to carbon generation. The Green Grid—a consortium of information technology companies and professionals seeking to improve energy efficiency in data centers around the globe—is developing the most advanced metrics for data center efficiency [11, 12], and the power management technology should support these reporting requirements.

SOLUTIONS: A POLICY-DRIVEN APPROACH TO POWER MANAGEMENT

A typical data center is managed by deploying management systems for monitoring the health of the computing infrastructure and performing various management functions including error detection, failure resolution, server provisioning, and service deployment to servers. We can extend this management infrastructure to enable the dynamic power management approach described above by introducing a new component: a server-resident PM that meets the requirements for power monitoring and control (see Figure 4).

The PM is responsible for enforcing power management directives that can be specified for each server. This is done by monitoring the appropriate power and thermal sensors on the servers, and by controlling the appropriate effectors that allow the PM to control the servers' power consumption as directed.

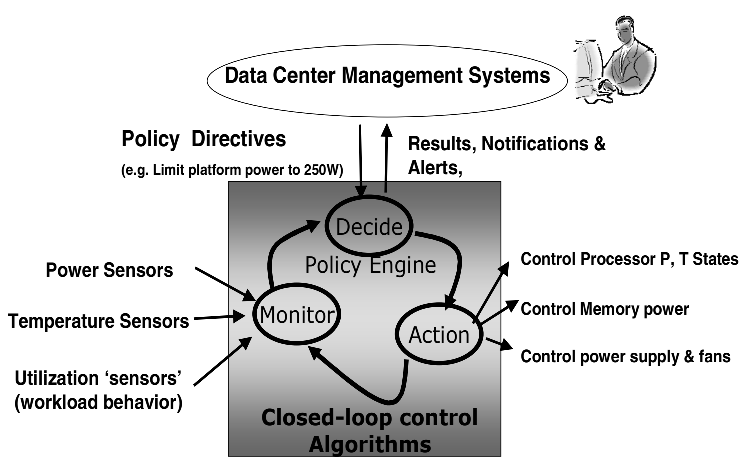

Figure 4: Power Management architecture

Intel's current series of server processors contains mechanisms that help control the power consumed by the processors. In addition, memory controllers provide ways of controlling power consumed by the memory modules. These can be used by the PM as low-level *effectors* for power control.

The interaction model between the PM and the data center management software is shown in Figure 4. The PM receives a policy directive from the Data Center Management System software. The policy directive may specify a power limit that is to be maintained for the server.

The PM has three major components operating in a closed loop manner: monitoring, policy engine, and control. The monitoring module is responsible for monitoring the sensors to determine if a new set of settings of the effectors needs to be enforced. This is done primarily by comparing the actual values of power and thermal against the policy limits. The policy engine decides a new set of controls based on the deviation of the actual power consumed from the limit and the utilization of various components. The control module then enforces the new settings as determined by the policy engine using mechanisms appropriate for that particular effector.

The sensors used by the monitoring module are the platform power sensors (including system and component powers) and the thermal sensors (system inlet/outlet temperature and component temperatures). Sensors providing information about the utilization of the components are used to determine the behavior of the workload.

The control module uses the effectors described earlier to effect power control on processor and memory. There are also additional effectors for components provided by the OEM (like power supply, fans, etc.) that are used by the PM to achieve greater power control.

When the power consumption exceeds the power limit, the PM will choose a set of settings for various effectors (processor, memory controller, etc.) that effectively reduce the power consumption of the platform. When power consumption goes below the limit, any restrictive controls previously placed on the processor, memory, and other effectors are relaxed. This is done in an iterative manner to account for the fact that the choice of effector settings may not have been accurate given the uncertainties in the behavior of the workload and the uncertainties in the parameters used by the PM to determine these settings. If the PM exhausts the use of all the controls at its disposal and still finds power consumption above the specified limit, it can generate alerts.

The PM exposes an abstracted interface for interacting with the external management software. It also requires interfaces for communicating with monitored components (e.g., temperature sensors, platform power measurements) and controlled subsystems (e.g., processor and memory modules).

The PM can be deployed in various ways:

- As a firmware running on a dedicated microcontroller in the server.
- As part of the baseboard management controller (BMC) that also performs other system management functions in the server.
- As an in-band agent in the operating system.
- As a combination of the above.

While the PM provides the capabilities described in the previous paragraphs, it depends on external management software to specify the policy parameters for it to operate effectively. The PM exposes an interface to the management software for this purpose. The interface includes commands to read power consumption and thermal data. The interface also allows management software to specify commands to set and get power control policies (e.g., set power limit) and to receive alerts from the PM.

The external interface is exposed as extensions to industry-standard server management protocols such as Intelligent Platform Management Interface (IPMI) [14] and Web Services Management (WS-Man) [15]. While IPMI is widely used in the industry today for server management, we anticipate WS-Man to gain more acceptances in the future.

CASE STUDIES AND EARLY RESULTS

To demonstrate the value of a policy-based dynamic power management approach using a platform-resident PM, we implemented the PM and conducted two sets of experiments: one set at a pilot data center, and the other in our internal labs. The rest of this section describes these experiments and the early results that demonstrated the value of this approach.

We conducted a proof of concept (PoC) at a top Internet portal customer's data center as a pilot project. The objective of this PoC is to maximize the number of servers allowed in a single rack within a given power envelope while maintaining maximum application performance.

Table 1 lists the use cases we developed for the PoC. At the beginning of test, we recorded the nameplate value of the servers (~350W) which the customer uses to populate their data centers today. We then installed servers with

PMs in the rack and measured the actual maximum power consumption of each server at peak search workload using the PM (~310W). We then used the observed maximum power value as the baseline for setting power limits for the servers and for determining number of servers per rack. We allowed a 10% headroom, to make sure that the power consumption at the rack level did not exceed the power envelope for prolonged period of time (10 min. or longer). The PM automatically adjusts power consumption toward this target, while continuing to deliver maximum performance for the given workload.

The data center management system communicates with the PM using IPMI [14] to continuously monitor actual power consumption of each server. It then aggregates power measurements at the rack level to make sure that the rack-level power envelope was not violated. The data center manager is used to set power limits dynamically for each server as desired to achieve the IT management policy. If the PM cannot maintain the limit set, or the data center manager observes a trend towards violating the rack-level power envelope, it resets the limits appropriately to ensure that the rack power envelop is not violated. With the interaction of the data center manager and the PM, the customer can safely achieve the maximum number of servers for a given rack-level power budget, thereby increasing the density of servers on a rack.

Table 1: Rack-level power optimization use cases

Use Cases	Description
Get power consumption on each server	Using the Policy Manager to dynamically gather point time power consumption from each server on the rack.
Estimate total power consumption of a rack	Estimate rack-level power consumption by summing up node-level power consumption; display on and notify to console as appropriate.
Optimize rack-level policy within a given power envelope and server workload	At rack level, analyze the power consumption of each server, overall power consumption, rack-level power envelope, and targeted performance goals (utilization, response time, query queue length, etc.) as well as other factors important to Baidu to determine the optimal power distribution policy. Baidu will set the policy and optimization strategy based on their work load and priority.
Set policy to servers on the rack	From the console, set policy to each rack in terms of particular power budget target that the server has to observe.
Node-level monitoring and tracking against policy	Leveraging Intel Node Server features to adjust server power consumption to the target set by the policy within 60 seconds and maintain at the target until further notice.
Node-level alert and notification	Use Policy Manager to detect and send alert when a server fails to reach policy target in 60 seconds or maintain the target during operation.
Alert handling and mitigation	Once an alert is received, the console needs to automatically decide on a course of action to mitigate the risk–ignore, set a new policy, or shut down the troubled server. …

The initial result from the PoC described above is summarized in Table 2. The result shows the performance measurement and power reduction observed for a single server in the PoC. The server has two Quad-Core Intel® Xeon® processors configured with 16 GB memory and a PM. The server was running actual search workload at the customer site in a near-production test environment.

It is interesting to note that when the workload is around 1,500 concurrent searches (above average workload) and the PM imposes a power cap at around 270W, the server CPU utilization and throughput virtually remain the same, i.e., ~67% and ~4.7ms per search respectively. This means that with a PM and proper power limit, we could save 40W from a server without performance loss when the CPU is not fully loaded. This is a 13% power reduction without performance loss. Under this circumstance, we could add one more server to a rack that is populated with 6 servers within the same power envelop.

It is important to understand that the value of PM is dependent on the actual application running on the server, the typical workload, and configuration of the server itself. For each combination of server and workload it is running, the user should determine the desired control points that reduces power consumption with minimal or no performance impact.

Table 2: PM Test Result on a Single Node

PM Setting	Platform Power Consumption	Workload	CPU Utilization	Search Time
No PM	310W	1,468 concurrent searches	67.81%	4.79ms
PM Power Capping 265 W	270W	1,514 concurrent searches	67.83%	4.69ms

In the second set of experiments which we conducted in our labs, we further explored the value of a PM by setting policies at different levels. We populated a rack with Intel® Bensley servers equipped with a Quad-Core Intel Xeon processor configured with a PM. We integrated servers under test with a management console, so that the management console could get real-time server power-consumption data and define policies to set a power-limiting target, while maintaining best possible performance at the limit.

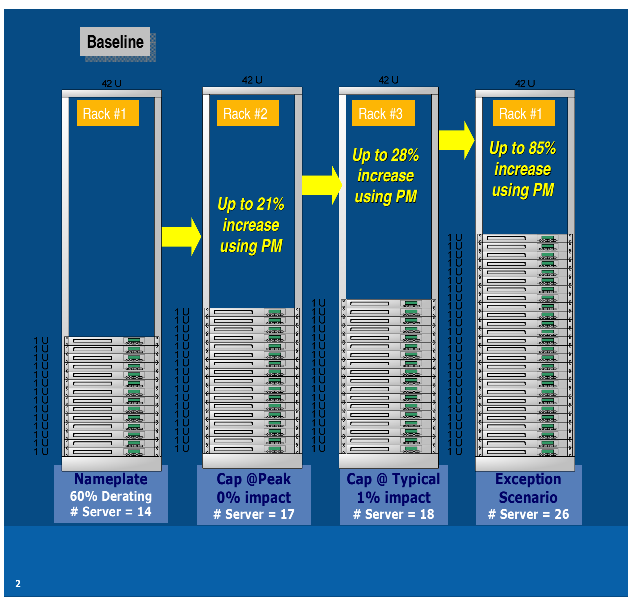

Figure 5: Policy Manager case study

Four different test scenarios were considered:

1. Populate the rack with nameplate power consumption (current customer practice).
2. Populate the rack based on power capping at maximum performance power measured for given search workloads (~280W).
3. Populate the rack based on power capping at ~98% of maximum power measured for given search workloads (~275W).
4. Populate the rack based on power capping at 90% of maximum power measured for given search workloads (~250W).

As shown in Figure 5, servers with a typical configuration, running representative workloads are populated in Rack #1. The rack is provisioned with a nameplate power with a derating factor of 60%.

Using the PM, as shown in Rack #2, by provisioning servers with power corresponding to maximum performance power, additional servers can be populated in a rack up to 21% more. Similarly, using the PM, Rack #3 shows an increase in server density of 28% when servers are provisioned with power that impacts peak performance 1% of the time. An exception scenario shown in Rack #4 is applicable when the power is severely constrained, for e.g., due to bad weather. In this case, the PM will limit power to the individual servers and hence to the whole data center, but it will allow the data center to operate in a stable environment. Each additional server defers data center capital expenditure by ~$2,000 [3].

Another example of the value of the PM is demonstrated in Figure 6. We show the power consumption and throughput measures with and without the PM for utilizations between 60% and 100% for WebBench load, which is a benchmark for Web traffic. For this workload, when the PM has a policy that limits the power to 20% below at 100% utilization, the impact to throughput is only 10%. The numbers are better at 80% utilization, where for a power reduction of 18%, the throughput is only 3%. It should be noted that typical servers in data centers run at utilizations well below 60% and as seen from the chart, the performance impact for a given power reduction is even less. Therefore, using the PM, we could safely craft power-capping policies to limit server platform power consumption with little impact on the peak performance of applications.

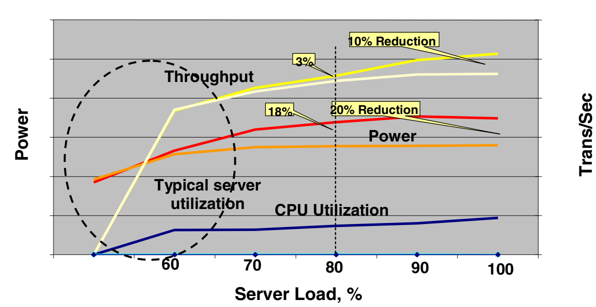

Figure 6: Power/performance with the Policy Manager

CONCLUSION

Current trends in data center power reveal a fundamental need for a power management capability on the platform that can be used to monitor power consumption and enforce power policies.

We have described an embedded policy manager (PM) as the foundational capability for a dynamic policy-based power management approach. The initial results of our implementation show that a dynamic policy-based power management approach using a PM can be used to increase server density within a rack power envelop, and to reduce power consumption with minimal performance impact.

In our future work, we will explore additional use cases and policies, and further investigate the benefits of fine-grained power-control methods.

ACKNOWLEDGMENTS

Special thanks to our colleagues Mike Patterson for providing regulatory requirements and Derek Collier for insightful suggestions during the review of this paper.

REFERENCES

[1] X. Fan, W. Weber, L. Barroso G; Bramlett, "Power Provisioning for a Warehouse-sized Computer," in *Proceedings of the ACM International Symposium on Computer Architecture,* June 2007, San Diego, CA.

[2] E. Castro-Leon, M. Chang, J. Hahn-Steichen, J. He, J. Hobbs, G. Yohanan, "*Service Orchestration of Intel-Based Platforms Under a Service-Oriented Infrastructure," Intel Technology Journal, Vol. 10, Issue 4*, November 9, 2006.

[3] J. Koomey, K. Brill, P. Turner, J. Stanley, B. Taylor, "A Simple Model for Determining True Total Cost of Ownership for Data Centers," *Uptime Institute White Paper*, October 2007.

[4] W. Felter, K. Rajamani, T. Keller, C. Rusu, "A Performance-Conserving Approach for Reducing Peak Power Consumption in Server Systems," in *Proceedings of the 19th annual international Conference on Supercomputing*, 2005, pps. 293–302.

[5] M. E. Femal, V. W. Freeh, "Safe Overprovisioning: Using Power Limits to Increase Aggregate Throughput," *Power-Aware Computer Systems*, Lecture Notes in Computer Science, Springer Berlin, Vol. 3471, 2005.

[6] J. Moore, J. Chase, K. Farkas, and P. Ranganathan, "Data Center Workload Monitoring, Analysis, and Emulation," *Eighth Workshop on Computer Architecture Evaluation using Commercial Workloads*, February 2005.

[7] R. Bianchini, R. Rajamony, "Power and energy management for server systems," *IEEE Computer magazine*, Nov. 2004, pps. 68–76.

[8] W. L. Bircher, L. K. John, "Power Phase Variation in a Commercial Server Workload," in *Proceedings of the 2006 international symposium on Low power electronics and design*, ACM, pps. 350–353.

[9] "Data Center thermal guidelines," at www.ashrae.org

[10] "Report to Congress on Server and Data Center Energy Efficiency," at http://www.energystar.gov/ia/products/downloads/EPA_Server_Report_Public_Draft_23Apr.pdf

[11] "The Green Grid Data Center Power Efficiency Metrics: PUE and DciE," at http://www.thegreengrid.org/gg_content/

[12] "Existing Metrics, Guidelines and Programs Affecting Data Center and IT Energy Efficiency," at http://www.thegreengrid.org/gg_content/

[13] PMBus standard information, at http://pmbus.info/

[14] Intelligent Platform Management Interface (IPMI), http://www.intel.com/design/servers/ipmi/spec.htm

[15] Web Services for Management (WS-Man), http://www.dmtf.org/standards/wsman/

AUTHORS' BIOGRAPHIES

David Filani is a Principal Engineer at Intel's Digital Enterprise Group. His current interests are platform power optimization and management, and he leads a team of architects focused on developing platform power management technologies. He holds a B.S. and M.S. degree in Computer Science from Arizona State University. He can be reached at david.filani at intel.com.

Jackson He is a Lead Architect in Intel's Digital Enterprise Group End user Platform Integration Group. He received PhD and MBA degrees from the University of Hawaii. His current focus is platform manageability and dynamic resource management in data centers. His e-mail is jackson.he at intel.com.

Sam Gao is a Program Manager at Intel's Digital Enterprise group in charge of directing server power management technology products. He received his MBA degree from the University of Chicago and an M.S. in Computer Science and Physics from Rutgers University. He is a graduate of Tsinghua University (Beijing, China) with a Bachelor's degree in Physics.

Murali Rajappa is a Platform Architect in Intel's Digital Enterprise Group Architecture and Planning Group. His

current focus is platform-level power management for next-generation computers with special emphasis on servers. His e-mail is muralidhar.rajappa at intel.com.

Anil Kumar is a Platform Architect in Intel's Digital Enterprise Group Architecture and Planning Group. His current focus is platform-level power management for next-generation computers with special emphasis on servers. He holds a PhD degree from the University of Connecticut and an MBA degree from Southern Methodist University. His e-mail is anil.k.kumar at intel.com.

Pinkesh Shah is a Platform Architect in the Digital Enterprise Group at Intel Corporation. He is involved in power management technology pathfinding for Intel's future server and client products. He received a PhD degree from the University of Arizona, an MSEE degree from Columbia University, and a BSEE degree from the New Jersey Institute of Technology. His e-mail is pinkesh.j.shah at intel.com.

Ram Nagappan is a Platform Power Architect in the Digital Enterprise Group at Intel Corporation where he is responsible for pathfinding for power and thermal solutions for Intel's next-generation server, client, and communications products. He received his Masters of Science degree in Engineering from Arizona State University. He can be reached at ramkumar.nagappan at intel.com.

THIS PAGE INTENTIONALLY LEFT BLANK

For further information visit:

developer.intel.com/technology/itj/index.htm